NETWORKS OF ASSOCIATION

TECHNICAL REPORT NO. 7

PREPARED UNDER
Contract Nonr 1305(00)

FOR THE OFFICE OF NAVAL RESEARCH

October
1954

DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.

TECHNICAL REPORT NO. 7
NETWORKS OF ASSOCIATION

The search for associations is essentially a search for ideas, although it will frequently be a preliminary step in the search for specific documents. Hence, the association "machine" must neither be required to perform as an index nor, more important, may it fail to show a relation between concepts simply because there is no single document exhibiting such a relation. Without a document about ABC, the relation between A, B, and C may have to be deduced from three documents dealing with AB, BC, and AC. The association system should, nevertheless, show that A, B, and C are all related; it remains for the index to tell for which combinations of these terms there are documents in the system.

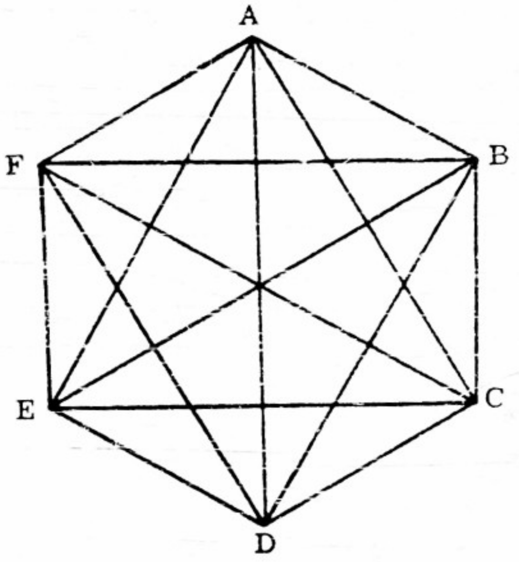
Since the existence of documents indexed by particular combinations is discoverable through the index and not the association machine, we believe that the mechanism of association should presently be limited to networks of two term associations. As pointed out in Technical Report No. 5[1] the listing of three-term combinations would require a device many times more complicated than that required to give two-term association. Whether

[1] "The Preparation of Manual Dictionaries of Association", Technical Report No. 5, Documentation Incorporated, April 1954, p. 5.

or not any of the additional effort is justifiable depends on whether the searcher would be aided by having the association machine report combinations of terms which have all been used together to index one or more reports, rather than combinations of terms which are closely related. This remains to be seen in practice. It might actually be a disadvantage to have an association system which would present A, B, and C only if a document indexed ABC exists.

The important function of the association machine is to present to the searcher for his selection those terms which are sufficiently closely related to each other to form a reasonable basis for a study or literature search in any desired combination. That this function can be performed without going beyond two-term associations can easily be demonstrated.

A searcher using the punched card system described in the previous report first selects the card representing a key term, A, in which he is interested. The positions punched on this card indicate the terms which are associated with it. He then selects any of these, say B, and superimposes the card for B with that of A. Since each term is coded in the same position on every card, the holes which now show through both cards indicate the terms which are associated with both A and B. This process of selection and superimposition can be continued as long as desired. The terms selected by this process form a network, any two of which are associated in some document. This can be represented graphically as follows, each chord indicating an association:



Such a network of two-term associations should certainly fulfill the conditions of the search. Clearly, these terms are all closely related (whether or not one document contains all of them). The mechanical display of such networks of association effectively solves the challenge of Vannevar Bush and provides the "coincidences" of ideas which Bernier called the most important characteristic of an information system.

Since we are here proposing that mechanical association of ideas is to be achieved by the superimposition of dedicated positions in a set of cards or plates, a mechanical dictionary of associations can be either a Batten system in which each term of the dictionary is a card or plate, or a system of language elements. Whether term cards or language element cards are used, the body of the cards will contain the same pattern of dedicated positions for all the terms in the system. The actual punching or use of a dedicated position will, of course, indicate an actual association in the system of the term designating the card (or term made up from a set of language elements cards) with the term punched on the card (or set of cards).

In the indexing machine, a hole common to two cards indicated a document as a member of the class which is the logical product of the classes designated by the two cards. A hole on the air card at position 475 and a hole on the ducts card at position 475 indicates that item 475 concerns air ducts. In the association machine a hole on the air card at position 475 and a hole on the ducts card at position 475 indicates that the term in the system which is numbered 475, say icing, is associated with air and with ducts, (A, D)*I. Note that in accordance with the analysis of the logic of association in Technical Report Number 4,[2] the association A*I. and A*D. does not tell us

[2] An Extension of the Algebra of Classes for the Association of Ideas," Technical Report No. 4, Documentation Incorporated, April, 1954.

whether or not there is in the system the association A *I *D. That is, we are not told by the association machine whether any particular item is a member of the logical product AID. (air· icing· ducts).

Although we are confident that the search of any system of information for networks of association should be some sort of a machine process, we must not lose sight of the fact that a manual dictionary with each term in the system denoted on a page (Cf Exhibits 1-3, Technical Report No. 5) will give us all associations of the terms in the system with any given term, just as a single Uniterm card will give us all the items which are members of the class denoted by the Uniterm. It is only when we wish to coordinate material on one page with material on another that the problem of mechanization becomes germane.

All the problems which were treated in the discussion of the indexing machine must also be considered with reference to the association machine, namely the number of cards or sheets; the number of dedicated positions, the percentage of use of dedicated positions, and the probability of false drops; and we will discuss each of these in turn. There is, however, an additional problem in the association machine which does not affect the indexing machine. The association machine must display the associated terms at every stage of the machines operation. The pattern of lights displayed by the indexing machine at the conclusion of a search represent numbers. These numbers can be reproduced automatically on a tape as the scanning frame on the indexing machine passes over the light dots on the screen.

In the operation of the association machine the selection of terms at successive steps in the associative process is made from the set of terms displayed at the prior step. This selection determines the associations displayed by the subsequent operations of the machine and is essentially a "feedback" device. But the very nature and purpose of the machine and the nature of the association process indicates that this feedback cannot be made to operate automatically.

Nothing is freer than the mind in making associations-anything can be associated by the mind with anything. The notorious fallibility of memory can be expressed as a tendency of the mind to forget associations previously made or experienced, or as a tendency to refer a newly created association to a past time. That is, the mind forgets observed associations and posits associations which never occurred.

The dictionary of associations corrects the fallibility of memory by presenting all associations in a system and only the associations in a system. But for any particular question or search put to a system of associations, certain of the associations may be irrelevant. This irrelevance is not a matter of logic but of purpose. So far as the dictionary is concerned one association is as good as another; but for a particular purpose motivating a particular search the purpose must guide the selection of terms constituting the network. Hence, as we have noted above, in order to make possible this purposive feedback, the associations, at every step of the machines operation, must be displayed to the searcher or operator.

The different methods of display form a group of problems which will be handled in a separate paper since, in the balance of this paper, we will be concerned only with those problems of the association machine which are analagous to the problems of the indexing machine.

In any system the actual associations will, of course, depend on the subject matter of eachdocument. We can, however, make some statistical calculations based on reasonable assumptions and as in previous reports, for this purpose we will assume a collection of 50,000 items, a dictionary of 5,000 terms, and the use of 10 terms to analyze or index each document.

If any word in our dictionary is used only once to index one document, it will be associated with only nine other terms. The card for such a term would thus have punches for at least nine other terms. If, however, we assume a uniform use of the terms in the dictionary, each term will be used in the analysis or indexing of 100 documents.

$$\frac{50,000 \text{ documents} \times 10 \text{ terms per document}}{5,000 \text{ terms}} = 100$$

In this uniform system, the maximum number of punches on a term card will be nine times the number of documents indexed by that term, or 900.

We shall assume that the indexing of a document is independent of any other documents, and that all subjects are equally likely. Then, the probability that term A is used to index a certain document is 1/500 (since A is used for 100 documents out of 50,000). If term B is independently chosen from the rest of the 4,999 terms, each equally probable, then the probability that both A and B are entered for one document is 1/500 x 9/4999, and the probability that they were not both used for that document, $1 - 1/500 \times 9/4999 = 1 - \frac{9}{2,499,500}$.

The probability that they are associated is the probability that they have been used at least once, or one minus the probability that they were not used for any of the 50,000 documents:

$$\text{Prob } (A{*}B) = 1-\left(1 - \frac{9}{2,499,500}\right)^{50,000} = 0.1648$$

We should, therefore, expect each term to be associated with an average of 50,000 x 0.1648 or 824 other terms.

It is undoubtedly not true, however, that if A has been used to index a particular document, the remaining 4,999 terms are equally probable choices for the other nine index entries. When one aspect of a subject is known, there are certain terms which are more likely to be needed to complete the description. Let us assume that a set of 999 terms (to be called M_A) are about 10 times as likely to be used with A. as the other 4,000. Then the probability that A is associated with B is 0.474 if B is one of the 999 in M_A, and 0.0622 if B is one of the 4,000. The expected number of paired associations per term, or punches per card is 722 for these conditions.

If we do not know in advance whether B is among the 999 or the 4,000, the probability of its being associated with A is simply 722/4999, or 0.144. The probability of an arbitrary third term C being associated with both A and B is $(0.144)^2$, or 0.0207. When the cards for A and B are superimposed, therefore, the most likely number of common holes is 104. Similarly, there will probably be just 15 terms associated with each of three others chosen at random, and only two with each of four terms.

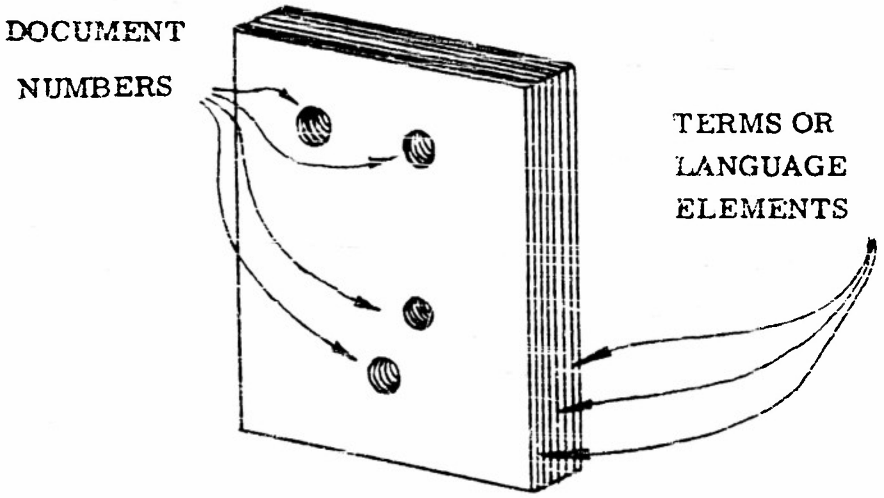
It is rather academic, however, to consider terms chosen at random. As described above, the second term examined will ordinarily be chosen from among those associated with the first; the third, from those associated with both of these; etc. The probability that C*B is greater if we know that A*B and C*A, than it would be if we did not have these facts, and we must take this into account.

Let us suppose that we have chosen B from among the terms associated with A, and now superimpose the two cards to see which terms are associated with both. The number will depend on the relationship between A and B. If more of the 999 terms of M_A are also frequently used with B, more terms are likely to be associated with both A and B. We shall assume that M_A and M_B have 499 terms in common. Of these, $(0.474)^2 \times 499$, or 112 will probably be associated with both A and B; of an additional 1,000 terms, 0.474×0.0622, or 29.5; and of the remaining 3,500, $(0.0622)^2$, or 13.5. We therefore expect 155 common holes when A and B are superimposed.

The next step is to choose one of these terms, C, and superimpose it with A and B. We shall assume that M_C also contains the 499 terms common to M_A and M_B; the remaining 500 not common to either. Then the probable number of terms associated with all three is $(0.474)^3 \times 499 + 0.474 \times (0.0622)^2 \times 1500 + (0.0622)^3 \times 3000$, or 57. Similarly, superimposition of four cards narrows the field to $(0.474)^4 \times 499$)t e second and third terms become negligible), or 25; five cards, to 12; and six cards, to six terms (5.7, on the average, not counting the holes for the six being superimposed; which will also be punched on each card).

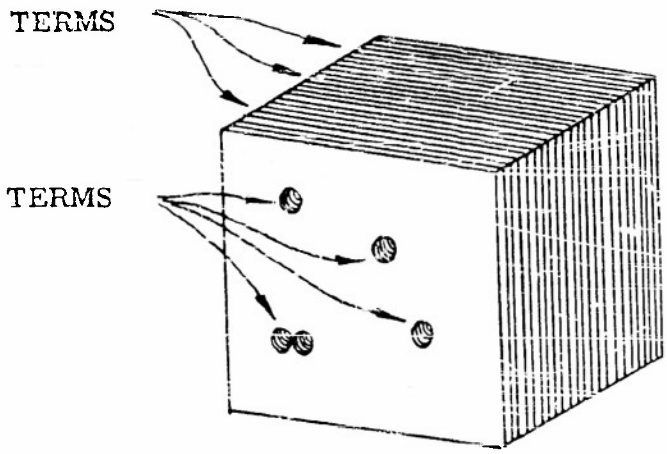
The terms on the six cards form a network of associations, every pair being associated, as described above. At any stage of association, the network can be modified or enlarged, by going back and changing the selection of terms to be superimposed.

It will be recalled that the indexing machine was, in effect, a three dimensional body of information, with one dimension (depth) representing the terms of the system, and the other two (height and width), representing the coordinates of any document number.



When the individual cards or sheets are language elements rather than terms, the depth of the solid (no. of terms) will be small as compared with the area of each sheet. In fact, the only restriction on reducing the number of sheets is the problem of superimposition. We cannot tolerate a situation in which more than half the holes on any sheet are punched, and we prefer having enough sheets to restrict the average density of punching to 1/3 the holes on any one sheet.

The association machine is similarly a three dimensional figure, which can be regarded as cube of information (even though it may not be a physical cube.) For in the association machine all dimensions measure terms in the system, and the depth of the solid, in terms, is always equal to the number of occupied dedicated positions.



The question we must now answer is this - should the cards or sheets in the association solid be language elements or terms?

With reference to the indexing machine, our decision to use language elements was based on a number of considerations, namely:

1. The low density of punching on a Batten card
2. The size of a sheet necessary for a large collection of documents.
3. The reduction of a system from 5,000 term cards to 500 language element cards would not increase the density of use of any card beyond tolerable densities.
4. The size of the sheet necessary to handle a large collection of documents made it desirable to eliminate the necessity for adding sheets for new terms.
5. The use of small term cards (Batten cards) would necessitate additional sets of term cards whenever the number of documents in the system exceeded the capacity of a card.
6. There would always be some addition of new terms to the system, or sets, and no way of telling in which set any particular term will be found. Suppose for example, a Batten system with 20,000 holes per card were used to catalog 100,000 items. There would be 5 sets of term cards but the sets would not be absolutely uniform since some terms would not be used in all sets. On any search, however, we would have to search all five sets.

When we turn from the indexing machine to the association machine the same series of considerations leads to a decision to use term cards instead of language element cards.

1. It appears from the above statistical considerations that the density of posting of associated terms on any term card will be high.
2. Since the size of the sheet necessary for a large system is determined by the number of terms in the system, and not by the number of documents, the association card can be relatively small.
3. The reduction of a system from 5, 000 term cards to 500 language element cards would increase the density of use of the cards beyond tolerable densities.
4. Since the sheets or cards are relatively small the addition of new sheets for new terms does not present any unusual difficulties.
5. Since the number of positions dedicated on any sheet would provide for all actual and potential terms in the system, we would never require a new set of term cards but only addition of individual term cards.